Selecting the best statistical distribution using multiple criteria

(Final version was accepted for publication by *Computers and Industrial Engineering*, 2007).

ABSTRACT

When selecting a statistical distribution to describe a set of data there are a number of criteria that can be used. Rather than select one of these criteria, we look at how multiple criteria can be combined to make the final selection. Two approaches have previously been presented in *Computers and Industrial Engineering.* We review these, and present a simpler method based on multiplicative aggregation. This has the advantage of being able to combine measures which are not measured on the same scale without having to use a normalisation procedure. Moreover, this method is scale-invariant, thus re-scaling the criteria values does not affect the final ranking. The method requires strictly positive criteria values measured on a ratio scale.
The proposed multiplicative method is transparent, simple to understand, apply and communicate.

1. Introduction

We consider the problem of selecting a probability distribution to represent a set of data. It has been pointed out that there are several criteria that can be considered when making this choice. The aim is to use all of these criteria when making the choice, rather than select one criterion for this purpose. To illustrate, Wang et al (2004) provided data from an engineering problem involving machine tools. They present five criteria:

- deviation in skewness and kurtosis,
- average deviation between the theoretical probability distribution function and the empirical one,
- average deviation between the theoretical cumulative distribution function and the empirical one,
- the Kolmogorov-Smirnov test statistic
- a subjective score (obtained from a group of experts in the field of study and statistics) “on the user friendliness of the distribution and the frequency of its use in the field, and the fitness of properties and characteristics of the distribution to the sampled data”.

Our focus in the present paper is on the method used for making the final choice of distribution. We shall not discuss whether the above are appropriate criteria – the analyst may wish to use a different set – it is the method for analysing the criteria values that is of interest here. This question has received some attention, and we shall look at what has been proposed. We shall consider the method of Wang et al (2004), then the critique of

this by Ramanathan (2005) and his proposed approach. Finally we present an extremely simple method which does not suffer from any of the disadvantages of either of these published methods.

2. The method of Wang, Yam and Zuo

Wang et al (2004) used a weighted sum of the criteria to compare distributions, with the weights provided by experts. They felt a need to standardize the criteria values to lie in the zero to one range, with higher values being preferred. Their way of achieving this is rather unexpected, for each criterion *v*, they use the following formula to convert to a standardized score: $r(v) = 1/(1 + cv^2)$ where c is a positive constant whose value is chosen to ensure the range is (0,1). Inspection of their table of standardized scores indicates that none of them have the maximum value of one. Instead, the transformation has been carried out with convenient round numbers for the constant, such as 10 and 500 – possibly from a linear search – so that the largest score exceeds 0.9 for each criterion. In fact, inspection of the above formula for $r(v)$ immediately indicates that a value of unity cannot be achieved for any real data, as this would require $cv^2 = 0$. Ramanathan (2005) criticised the arbitrary nature of the choice of the constants, and by selecting a different value for one of them was able to demonstrate that a different distribution comes out as the best.

Quite apart from the choice of constants, there is the wider question of the functional form of the transformation. One possible reason for wanting to standardize is that one does not want one criterion to dominate purely by virtue of the fact that its typical values happen to be greater. For example, the mean value for the deviation in skewness and kurtosis was 2.2 whereas the mean value for the deviation from the cumulative distribution was 0.0009. Another reason for standardizing is that people find it easier to work with numbers which all lie in the same convenient range. A third reason arises where different criteria are measured in different units and one wants to make them dimensionless so that they can be added together. The trouble is that there are numerous ways of standardizing – how does one choose? This seems to depend on one's field of practice; for example statisticians are used to converting data to z-scores, whilst operations researchers use other approaches, each will have merits and drawbacks. The fact is that there is no clear best way. Different transformations will, in general, not lead to the same final outcome. Pomerol and Barba-Romero (2000) illustrate this with a simple example, emphasising that "prior normalization of the data is not a neutral operation, and the final result of aggregation may well depend on the normalization method used". This is a crucial point, and one that is not widely appreciated. To overcome this obstacle we shall demonstrate an approach which does not require any standardization.

3. Data envelopment analysis

Ramanathan (2005) proposed the use of data envelopment analysis (DEA) to select the best distribution. The distinctive feature of this approach is that the weights are not decided by experts, indeed it does *not* employ a common set of weights to apply to all the candidates/alternatives. Instead, separate sets of weights are calculated for each candidate using a constrained optimization procedure. The aim of the optimization is to choose weights which show each candidate in the best possible light. The score for *each* candidate distribution is computed in turn as follows:
The score is a linear combination of the raw performance data, with the weights chosen so as to optimize this candidate's score subject to constraints which ensure that when these weights are applied to all other candidates none of the scores exceed 100%. Any candidate with a score of 100% is labelled efficient; the other candidates are inefficient and are no longer considered. They are rejected because despite having such freedom in choosing weights, they were unable to achieve the best score. Computationally, the score for each candidate is obtained by solving a linear programming problem.

We do not feel that DEA is an appropriate technique for this selection problem for the following reasons.

a) <u>DEA is not designed for selecting a single winner</u>.
   The original intention behind DEA is to identify the best-practice efficient frontier, not a single candidate. This frontier is piece-wise linear and consists of the candidates deemed efficient together with the linear facets that join adjacent ones together. Hence, it is not normal to have a single best candidate in DEA. Ramanathan (2005) in his DEA analysis used all of the criteria listed previously apart from the subjective one, he found that two of the six distributions considered were efficient for the machine tool data, namely gamma and Weibull. He therefore had to apply further analysis to make the final selection. Had more distributions been considered from the outset, it is likely this shortlist would have been larger than two, for it is well established in DEA that having more candidates leads to more of them appearing on the frontier.

b) <u>DEA may completely ignore the weaknesses of some candidates</u>.
   Consider the case where all the criteria are of the type where lower values are preferred. Any candidate which has the best score on one criterion will automatically qualify as efficient – irrespective of how poorly it scores on all the other criteria! This arises from the flexibility in allocating the weights – DEA seeks the optimal score for each candidate in turn, and this can occur by focusing on the best criterion and completely ignoring the rest! This is clearly no way to proceed in the selection of a statistical distribution.

   In the more general case where there are also output criteria (those where higher values are preferred), a similar issue arises. This time, whichever candidate has the best score on any ratio between one output and one input will automatically be deemed efficient – irrespective of how poorly it scores on all the other criteria. If there are m input criteria and n output criteria, then there are mn different possible

ratios and so up to *mn* candidates can be rated efficient in this way.

c) DEA becomes *less* discriminating as *more* information is provided.
As already indicated above, an unfortunate feature of standard DEA is that as the number of criteria is increased, the number of efficient units will also tend to increase. This is because the introduction of more variables means that a candidate is more likely to have something that it scores well at, and the weights will be chosen to emphasise this aspect so as to improve its overall score. Thus, the *more* criteria for which we have data, the *less* discriminating the method becomes!

There is an extensive literature of variations to DEA which presents techniques to improve discrimination. These include numerous ways to impose restrictions on the weights (see the survey by Allen et al, 1997), aggregating scores from multiple sets of weights (called cross-efficiency, see Doyle and Green, 1995), and removing constraints to allow scores to have a greater range (called super-efficiency Andersen and Petersen, 1993). Ramanathan opted for the super-efficiency approach, which led to the gamma distribution being chosen – this was also the one selected by Wang et al. Adler et al (2002) have reviewed ranking methods in the DEA context. These DEA variations are designed to make a further selection from the frontier points. The trouble is that there exist many such DEA extensions. Indeed Adler et al (2002) reviewed sixteen of them (including super-efficiency) classified into 6 groups; they conclude 'whilst each technique may be useful in a specific area, no one methodology can be described here as the panacea'. Thus it is not obvious how to proceed if DEA is used. The approach we present here avoids such complications and is far easier to execute and understand.

4. A simpler method

The method we propose is strikingly simple, yet has clear advantages over the approaches described so far. For the case where all criteria are of the lower-is-better type, we simply multiply together the criteria scores for each candidate to obtain its overall score. The candidate with the lowest overall score wins.

Our criteria consist of deviations between theoretical distributions and observed ones. A score of zero corresponds to a perfect fit i.e. no deviations between the theoretical distribution and the observed – obviously this is unrealistic. Nevertheless, in general we should exclude the possibility of zero scores on one criterion combined with non-zero scores on another, because a zero causes the overall score to collapse to zero irrespective of the ratings on the other criteria. Note that we cannot overcome this by adding a constant to each value to shift away from zero, this is because this would destroy the ratio scale property (i.e. equal variations corresponding to equal proportional variations). We require ratio scale measurements to provide us with scale invariance. Thus if we included a subjective criterion, the people doing the rating would have to be directed not to choose

zero ratings.

If we compare this multiplicative approach with the method of Wang et al (2004), we observe that one fundamental advantage is that there no need to choose a normalisation procedure for the data. This is because one no longer has the incommensurability problem – we are no longer *adding* together quantities measured in different units, so there is no need to make them dimensionless. Variables measured in different units can be multiplied directly. Also, rescaling any variable has no effect on the outcome because the overall product score will be affected by the same factor for each candidate. For example, if all values on a particular criterion are multiplied by 100 this does not affect their relative ranking as the overall scores remain proportional to what they were previously. Thus criteria which naturally have smaller values will not dominate on this account alone. However, we are not permitted to add a constant to the values of any variable, as this would destroy this proportional scaling property. It follows that we have to use data that is measured on the ratio scale. The four non-subjective criteria listed above have a clear absolute zero, corresponding to a perfect fit to the data, and so would seem to satisfy this requirement.

If one wishes to attach 'weights of importance', one does this using exponents i.e. by raising each criterion to a power. Note that such exponents are not equivalent to, and do not have the same interpretation as weights used in a weighted sum. For a weighted sum the (relative) weights represent the 'substitution or exchange rate between criteria', e.g. for the score function $2v_1 + v_2$ an increase of one unit in criterion v_1 can be offset by a drop of 2 units in criterion v_2 , leaving the final score unchanged. When exponents are used, this exchange is not between units but between percentage changes in the criteria. Thus if a criterion is given an exponent weight of p in the scoring function S, then a marginal change of, say 1%, in this criterion will lead to a p% change in the score. In economics p is referred to as the *elasticity* of S with respect to changes in the criterion; it is the percentage change in the score S, divided by the associated (marginal) percentage change in the criterion.

If we have 'more-is-better' criteria as well as 'lower-is-better', then we take separate products for of each of these two types and divide one by the other. If we place the 'more is better' product in the numerator, then our overall score function will be of the form where higher values are preferred.

Let us now consider the criticisms that we made above for DEA, to see if they apply to this method. Firstly, we said DEA was not suited to picking a single winner because typically it resulted in several candidates getting top score. With the score function that we are proposing it is possible that two or more units may end up with the same score, but this is unlikely. Secondly, all criteria in the analysis are taken into account. There is no way of hiding poor attribute levels. Thirdly, this method does not become less discriminating as more criteria data are used. In fact, it probably becomes more discriminating: if one starts with a single criterion there may be ties for first place, but as more criteria are added this becomes less likely.

5. Comparison of results

Since the numerical results from the various methods described are not comparable, we shall instead look at the rankings of the candidates. Wang et al.(2004), using the five criteria listed in section 1, selected the gamma distribution as the best, with the Weibull as second best, and the normal distribution as the worst. The DEA analysis of Ramanathan (2005) excluded the subjective criterion, but nevertheless ranked these same distributions in first, second and last places.

The simple multiplicative approach of the present paper (without weights) also ranked the gamma distribution as the best, the Weibull as second best, and the normal distribution as the worst. All of the ranks using this method remained identical whether or not the subjective criterion was included. A comparison of the results arising from the three methods appears in Table 1.

Table 1
Ranking of statistical distributions based on three different methods.

Method	Beta	Gamma	Weibull	Lognormal	Normal	Extreme-value
Wang et al.	4	1	2	3	6	5
DEA	5	1	2	4	6	3
Multiplicative	5	1	2	3	6	4

We also tried the multiplicative approach using the weights of importance as supplied by experts in Wang et al. (2004). These were not intended for use as exponents, but doing so remarkably resulted in identical rankings to those of Wang et al.

6. Conclusion

What are we to conclude from these comparisons? We see that there is no great disagreement between the methods when ranking the candidates. Whilst this may be reassuring, a cynic might then argue that it does not matter which method is used for making the choice. We however would argue that the method we have presented is preferable because it has a number of advantages. First of all its simplicity makes it very transparent and easy to apply and communicate to others. Secondly, the multiplicative form avoids the commensurability problem which arises when trying to add together criteria measured in different units. Thirdly, the fact that some criteria are numerically much greater than others by virtue of their units of measurement makes no difference, there is thus no need to normalise the data and so one is not concerned with deciding which normalisation transformation should be used.
The proposed method is applicable for strictly positive criteria measured on a ratio scale.

REFERENCES

Allen R., Athanassopoulos A., Dyson R.G., Thanassoulis E. (1997). Weights restrictions and value judgements in data envelopment analysis: Evolution, development and future directions. *Annals of Operations Research* 73: 13-34.

Adler, N., Friedman, L., and Sinuany-Stern, Z. (2002) Review of ranking methods in the DEA context. *Eur. J of OR* 140, 249-262.

Andersen P., Petersen N. C. (1993). A procedure of ranking efficient units in data envelopment analysis. *Management Science* 39(10), 1261-1264.

Doyle JR, Green RH. (1995). Cross-evaluation in DEA - improving discrimination among decision making units. *INFORS* 33:205-222.

Pomerol, J-C, and Barba-Romero, S (2000). *Multicriterion Decision in Management – Principles and Practice*. Kluwer, Boston.

Ramanathan, R. (2005). Selecting the best statistical distribution – a comment and a suggestion on multi-criteria evaluation. *Computers and Industrial Engineering*, 49, 625-628.

Wang, Y., Yam, R.C.M., and Zuo, M.J. (2004). A multi-criterion evaluation approach to selection of the best statistical distribution. *Computers and Industrial Engineering*, 47, 165-180.